AN INFORMATION-RICH CHARACTER WEIGHTING PROCEDURE FOR PARSIMONY ANALYSIS

A.G. RODRIGO

Department of Zoology, University of Canterbury, Christchurch 1, New Zealand.

(Received 2 May, 1989; revised and accepted 8 June, 1989)

ABSTRACT

Rodrigo, A.G. (1989). An information-rich character weighting procedure for parsimony analysis. *New Zealand Natural Sciences* 16: 97-103.

A weighting procedure is proposed which takes account of prior information pertaining to the characters used in a parsimony analysis. This information comes from specific knowledge about the biology of the group in question, as well as general evolutionary theory. The weighting procedure consists of three stages: (1) an initial parsimony analysis followed by (2) an examination of the character consistency indices and associated character weights, with reassignment of weights based on prior knowledge of the group; and (3) a reanalysis using the weighted data. The procedure is an iterative one, and can be terminated once the resultant tree has converged to a "constant value", or after a predetermined number of runs. The resultant tree may or may not be as short as the most parsimonious tree. It is argued that in taking account of prior information, the proposed procedure is information-rich (IR). Finally, the procedure is shown to be one of a family of IR techniques which are commonly used in parsimony analysis.

KEYWORDS: information-rich - character weighting - parsimony - phylogeny.

INTRODUCTION

The application of character weighting procedures in taxonomic analysis has always been a contentious issue, particularly for phylogenetic systematists. Systematists try to remove personal bias from their taxonomies by developing "objective" methods of classification. However, every systematist accepts that there are always some characters which are less "reliable" as indicators of phylogenetic relationships than others. Convergent characters may evolve in distantly related groups, either as a result of similar environmental pressures, or random genetic drift. Characters may also be misclassified through some error of interpretation on the part of the taxonomist. It seems clear that for any analysis which attempts to determine the phylogenetic relationships between groups of organisms, these characters should be given a low weight relative to those which are good indicators of ancestor-descendant relationships. However, the realisation that this must be so does not make the task any easier. Two problems arise:

1) how can these characters be identified; and

2) how can character weights be assigned to these and other characters, to reflect their *relative* phylogenetic information content.

Most systematists agree that procedures for character weighting, while essential, should rest on objective foundations. As a result, *extrinsic* character weighting procedures (i.e., those which use information not obtainable from the matrix of character states and taxa in question) have been rejected in favour of *intrinsic* methods which are more "algorithmic" and less susceptible to personal bias (see the methods developed in Farris (1969) and Penny & Hendy (1985)). Extrinsic weighting procedures are a special class of *a priori* weighting methods (*sensu* Neff 1986). By definition, extrinsic information precludes the

use of consistency indices, and compatibilities, both of which are obtainable from the character-taxa matrix, and are therefore items of intrinsic information. In this paper I will use the terms "prior information" and "extrinsic information" interchangeably.

The reason given for rejecting extrinsic weighting is that there is seldom any information available to determine which characters are good indicators of phylogeny in the group being studied. This is only partially true: while we cannot assign absolute weights (i.e., interval or rational values) to all characters, there is always some qualitative information available on the relative value of some characters in the data set. This information can be elicited from research on the comparative biology of the taxa in question, as well as from a general theoretical framework of population and evolutionary biology. So-called "objective" methods do not incorporate such information, and proponents of these methods are prepared to sacrifice prior information for objectivity.

In this paper, a method is presented which takes account of prior information while at the same time preserving the objectivity of intrinsic techniques. For this reason, the method is called an information-rich (IR) weighting procedure.

As a method, IR weighting is primarily an algorithmic extension of the principles discussed by Neff (1986) (and anticipated by Hecht & Edwards (1976)) in relation to *a priori* character weighting. Furthermore, I will argue that it is, in fact, one of a family of procedures which are commonly used in phylogenetic analysis.

I have applied IR weighting with parsimony analysis, but the method is general enough to be applied to all phylogenetic procedures with only minor modification.

TERMINOLOGY

Phylogenetic analysis attempts to uncover the evolutionary relationships between groups of study organisms or *evolutionary units* (EUs). These relationships are often displayed as a branching diagram known as a *phylogenetic tree* or *cladogram*.

For each EU, systematists have at their disposal information pertaining to the characters which may be used to identify the EU. Care must be taken to distinguish between *characters* and *character states*: character states refer to the "values" of a particular character, e.g., the character "hair-colour" has "brown","black", "blond", and "red" as its character states. For computational purposes, then, each EU may be represented as a set of character states. The number of possible character state changes is known as the *range* of a character. The range of a character is equal to the number of character states minus 1. For any given tree, the number of character state changes per character is known as the *length* of the character. The ratio of range to length is known as the *character consistency index.*

The problem of phylogenetic analysis can be stated thus:

Given what is known about the evolutionary process, how can EUs and character state changes be assigned to the terminal nodes and branches of a cladogram, respectively, to project a scientifically acceptable hypothesis of evolutionary history?

A number of phylogenetic methods have been developed, the most popular of which is *parsimony analysis.* Parsimony attempts to find the tree which has the fewest character state changes. The most parsimonious, or minimal-length, tree is taken as a hypothesis of evolutionary history. (Cladists argue that parsimony is based on a philosophically sound principle: the best hypothesis requires the fewest assumptions. Farris (1983), for instance, equates "phylogenetic tree" with "hypothesis" and "character state changes" with "assumptions". Hence, it follows that minimising character state changes on a phylogenetic tree is equivalent to choosing the best scientific hypothesis. In the last section, I will argue that this is not necessarily true).

METHOD

IR weighting is a three-stage process:

1) A parsimony analysis is conducted, *without* weighting.

2) Characters are ranked on the basis of their consistency indices. The user examines the ranks of these characters, and changes those which conflict with prior information. As stated earlier, this information may take the form of biological principles, theoretical considerations, ontological and genetic evidence, as well as the shared expec-

tations of other systematists working on the same group of organisms.

3) A weighting criterion is applied, using the revised ranks (and the consistency indices corresponding to these ranks), and the analysis is repeated. This process continues until the resulting tree converges to some stable value, or after a predetermined number of iterations.

Each of these stages is discussed in more detail in the following section, and will be illustrated with reference to the hypothetical data set of a group of potentially interbreeding but geographically isolated sub-species of parasitic flukes and their character sets, given in Table 1.

AN ILLUSTRATIVE EXAMPLE

Stage 1

A parsimony analysis is conducted using the data. A number of computer packages are available for this analysis (e.g., PAUP (Swofford 1985) and PHYLIP (Felsenstein 1987)). The output of the analysis should include the number of hypothesised changes for each character. From this, we can calculate the consistency index of the ith character, c_i

$c_i = r_i / l_i$

where r_i is the range of character i, and

l_i is the number of hypothesised changes of i (i.e., its length).

For the hypothetical data, parsimony analysis results in the tree shown in Fig. 1a.

This stage is no different from any other intrinsic weighting procedure, in that it involves an initial exploratory analysis.

Stage 2

The character consistency indices are ranked in descending order, i.e., the highest consistency index is given a value of 1, the next highest, a value of 2, etc. In Table 2, these ranks are given in column 5.

Once these ranks are available, the systematist is able to examine the *hypothesised relative stability* of the characters, and reassign ranks in accordance with what prior information is available. For instance, in the example, we see that Character 8 (follicular or whole testes) is hypothesised to have changed more often than most other characters in the group. However, it can be argued that changes in testicular morphology can lead to dramatic changes in reproductive biology, which in turn lead to reproductive isolation. Since the group is known to be at least potentially interbreeding (bearing in mind that the group in question is a hypothetical one), it seems likely that reproductive characters will, for the most part, be highly conservative. The same can also be said for Character 6 (genital opening, left or right). Certainly, biological theory would suggest that these characters are probably more conservative than characters related to the assimilatory system (Characters 4 and 5).

On this basis, it would be justified to reassign the ranks of characters 6 and 8 to the highest

Character			Taxa S1	S2	S3	S4	S5	S6	S7
1.	Body shape	(1 = elongate; 0 = elliptical)	1	0	1	0	1	0	0
2.	Body size	(1 = <5 mm; 0 = >5 mm)	0	1	1	0	0	1	0
3.	Head collar	(1 = present; 0 = absent)	1	1	1	0	1	0	0
4.	Oral sucker	(1 = terminal; 0 = sub-terminal)	0	1	0	1	0	1	0
5.	Gut caeca	(1 = diverticulate; 0 = smooth)	1	0	0	0	1	1	0
6.	Genital opening	(1 = left; 0 = right)	0	0	1	1	0	1	0
7.	Testicular fields	(1 = anterior; 0 = posterior)	1	0	0	1	0	1	0
8.	Testes	(1 = follicular; 0 = whole)	0	1	0	0	1	1	0
9.	Testes	(1 = lobed; 0 = complete)	1	0	0	1	1	0	0
10.	Eggs	(1 = with filaments; 0 = without)	1	0	1	0	0	1	0

Table 1. Hypothetical character-taxa matrix consisting of 6 subspecies (S1-S6) and 1 hypothetical ancestor (S7), and 10 characters of a group of parasitic flukes. The hypothetical ancestor serves to determine the evolutionary direction of the characters.

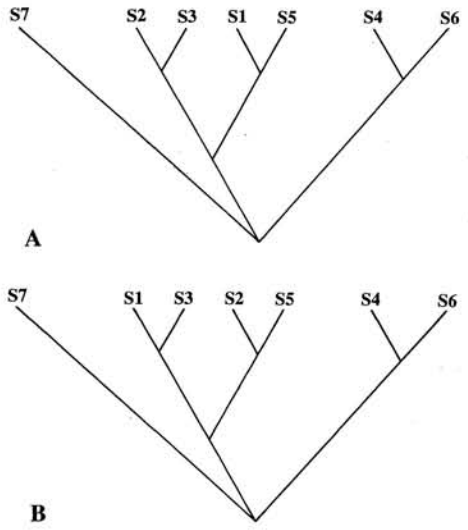

Figure 1. Phylogenetic trees derived using (A) unweighted and (B) weighted characters. The position of S1 and S2 differs in the two trees. (For consistency, the hypothetical ancestor, S7, has been positioned on a separate branch).

value, i.e., the rank of 1. Correspondingly, "new" consistency indices can be assigned to characters 6 and 8; in this case, the consistency index associated with rank 1 is 1.000. The new ranks are given in Column 6 of Table 2.

In essence, this stage involves the incorporation of information other than raw morphological data into the analysis. In practice, the taxonomist must be prepared to justify the reassignment of ranks, and the information which prompts such reassignment.

Stage 3

Consistency indices are reassigned in conjunction with rank reassignment because many weighting measures are functions of these indices. In this example, Farris' (1969) concave unbound-ed weighting function will be used. For the *i*th character, the weight, w_i is given by

$w_i = ((2n\text{-}3)c_i)^3\text{-}1$.

where n is the number of EUs.

These weights, applied to the reassigned consistency indices, are given in Column 7 of Table 2.

Stage 1 (a parsimony analysis) is repeated, this time using the weights given. The resulting tree is displayed in Fig. 1b, and the new consistency indices, and ranks, are given in Table 3.

Characters 6 and 8 now have ranks of 2. Clearly, this is more satisfactory than the previous scale, for the reasons mentioned above.

At this point, it is important to note that the unweighted length of this tree (i.e., the number of character changes, not corrected for weights) is one more than that of the tree derived in the initial parsimony analysis: the weighted tree is of length 22, while that of the unweighted tree is of length 21. Weighting has resulted in a tree which is not equivalent topologically to the most parsimonious tree, nor does it have the property of being a minimal-length tree (Fig. 1b). The consequences of this, and its justification, will be discussed in the next section.

The analysis is repeated, and the rank of Characters 6 and 8 are reset to 1, while all others

Character	Range	Length	Consistency index	Rank	New rank	Weight
1. Body shape	1	2	0.5	2	-	165
2. Body size	1	2	0.5	2	-	165
3. Head collar	1	1	1.0	1	-	1330
4. Oral sucker	1	2	0.5	2	-	165
5. Gut caeca	1	2	0.5	2	-	165
6. Genital opening	1	2	0.5	2	1	1330
7. Testicular fields	1	2	0.5	2	-	165
8. Testes	1	3	0.3	3	1	1330
9. Testes	1	2	0.5	2	-	165
10. Eggs	1	3	0.3	3	-	34

Table 2. Character consistency indicies, ranks, and weights derived from an initial parsimony analysis.

Character	Consistency index	Rank
1. Body shape	0.5	2
2. Body size	0.3	3
3. Head collar	1.0	1
4. Oral sucker	0.5	2
5. Caeca	0.3	3
6. Genital opening	0.5	2
7. Testicular fields	0.5	2
8. Testes	0.5	2
9. Testes	0.3	3
10. Eggs	0.5	2

Table 3. Character consistancy indicies and ranks after weighting.

adopt the new values of the weighted analysis.

Again Farris' weighting function is applied, and a parsimony analysis is conducted. The resulting tree, however, remains the same as that given in Fig. 1b. Similarly, the consistency indices of the different characters are the same as those given in Table 3. The analysis has "converged" to a single tree. This tree has desirable properties: the assignment of character state changes accords well with what is known about the biology of the group, and while it is not a minimal-length tree, it is only one unit longer.

DISCUSSION

To stress what was stated earlier, the systematist encounters two problems when attempting to weight characters for a phylogenetic analysis. The first of these concerns the differentiation of characters with a high phylogenetic information content from those with a low content. The problem is exacerbated by the fact that while we may have some knowledge about some characters, rarely do we have this kind of information about all characters.

The second problem is related to the first: how can a systematist assign weights to all characters when a) the appropriate weighting scale is unknown; and b) the phylogenetic content of only some characters is known (or can be guessed at).

The IR weighting procedure provides a solution to both these problems. First, it circumvents having to identify the phylogenetic information content of every character. Instead, by reranking the characters *after* an initial phylogenetic analysis, the systematist is free to decide on the *relative* reliability of only those characters for which there is any extrinsic information. The procedure therefore allows the initial analysis to determine the weights of those characters for which there is no information. Furthermore, decisions about relative stability (and consequently, relative weights) of characters are easier to make. It is easy to say, for example, that hair colour is less conservative than limb morphology, and at least as conservative as skin colour. It is more difficult, however, to assign an absolute weight to any of these features prior to an initial exploratory analysis.

Second, IR weighting frees the systematist from the task of selecting an appropriate weighting scale. Instead, all the systematist has to do is select one of a number of available weighting functions. Once this has been done, the reranking procedure will assign the appropriate weights to the characters. By reranking a character, IR weighting assigns a new consistency index to it. By doing this, a systematist is effectively stating the belief that the character can change as often as another with the same rank.

The scale of the weights is constrained by the choice of the weighting function. In the example given above, Farris' weighting function was used. However, there are a number of other functions available (Felsenstein 1981, Penny & Hendy 1985, Moody & O'Nolan 1987).

At this point, it should be noted that IR weighting can be used either as an exploratory procedure, or as a means of deriving a suitable tree. As an exploratory tool, IR weighting allows the user to compare the absolute length (as opposed to the weighted length) of the resultant tree with that of the tree prior to weighting. The absolute length of the weighted tree may be as short as, or even shorter than that of the unweighted tree. This is particularly useful when dealing with a large number of taxa (e.g., more than 20 EUs). This is because the procedures for obtaining the shortest possible tree become more cost-prohibitive as the number of taxa increases, and many computer packages resort to "best-approximation" methods.

Alternatively, a systematist may decide to

accept the weighted tree as the best hypothesis of evolutionary history, even though it is not the shortest tree. As in the example above, the weighted tree is considered to be a better hypothesis of evolutionary history because it incorporates more information about the characters than the unweighted tree does, at a "cost" of only 1 extra character state change. But can we justify not selecting the shortest tree as the best hypothesis of phylogeny ? What about Occam's Razor ?

At this point, it is worth reviewing the fundamental philosophy of parsimony analysis. When systematists use parsimony to construct hypotheses of evolutionary relationships it is rarely because they believe that evolution is parsimonious, i.e., that it proceeds with such a slow rate that all characters behave conservatively (Kluge 1984). Instead, parsimony is treated as a methodological tool, and as a way of constructing a hypothesis in a rational manner. Occam's Razor - "What can be explained by the assumption of fewer things is vainly explained by the assumption of more things" (Boehner 1957, translated by Kluge 1984) - is often cited as the fundamental motivation for the principle of parsimony in systematics. As stated earlier, cladists maintain that the minimal-length tree makes the least number of ad hoc assumptions regarding the multiplicity of character state changes.

This procedure is sound if there is no information about the nature of the characters selected. However, if information pertaining to the "conservativeness" of the characters is available from ontogeny, genetics or evolutionary theory, for example, then this procedure may falter. Consider, for instance, two characters, *a* and *b* for which there is a great deal of theory that indicates that the former is, in general, more conservative than the latter. However, after conducting a parsimony analysis, a systematist finds that, in the resulting tree, *a* has 3 changes while *b* has 1. While this may be the shortest tree for this data set, with the least number of ad hoc hypotheses, the character assignments it postulates is at odds with other theoretical considerations. To accept this tree would be to suggest that there exists an exception to the theory. If we accept that scientific theories are networks of hypotheses, theories, and observations, with each new theory or observation either supporting or casting doubts on others, it is important to realise that while we may have minimised the number of ad hoc assumptions for the tree itself, we have added one to the general body of biological theory. While it is true that exceptions abound in biology, many systematists would baulk at proposing such exceptions to the theory on the basis of what is really a hypothesis whose approximation to the truth is unknown (even, unknowable ?). A better tree would be one which preserved all relevant information, even at the cost of some units of length.

Finally, it should be noted that the weighting criterion proposed here is one of a family of information-rich procedures. Others in this set of procedures include Dollo parsimony (which has been formalised as a tree reconstruction procedure by Farris 1977), and the outgroup analysis of the polarity of character states (Watrous & Wheeler 1981).

Dollo's Law states that there is a smaller likelihood that complex structures would arise convergently, compared to simple structures. Dollo parsimony incorporates this by allowing only one forward change, while optimising the number of reversals. In Dollo parsimony, this information about the nature of character state change is supported by a background of evolutionary theory.

Outgroup analysis is a method by which ancestral character states may be determined by recourse to the distribution of these states in groups which are closely allied to the subject EU. It is argued that character states which are present in both outgroup and ingroup are likely to have been present in the ancestor of both groups. This information (which is not present in the EU-character matrix) allows the construction of a rooted tree, i.e., a tree which is not just a hypothesis of evolutionary relationships, but of evolutionary history.

I will conclude by noting that the techniques which are currently available for phylogenetic analysis are constantly being revised and enhanced so as to develop a family of procedures which take account of the diverse sources of information from which systematists must draw their conclusions. Information-rich procedures must be developed, but in such a way that these

methods are in harmony with intrinsic character weighting methods.

ACKNOWLEDGEMENTS

Thanks must go to Peter Johns and Richard Holdaway for their constructive comments on the first draft of this paper. Phil Garnock-Jones and another anonymous reviewer provided useful comments which allowed the manuscript to be further improved.

REFERENCES

Boehner, P. (1957). *Ockham - philosophical writings*. Bobbs-Merrill. Indianapolis.

Farris, J.S. (1969). A successive approximations approach to character weighting. *Systematic Zoology* 16: 44-51.

Farris, J.S. (1977). Parsimony under Dollo's Law. *Systematic Zoology* 26: 77-88.

Felsenstein, J. (1981). A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biological Journal of the Linnean Society* 16: 183-196.

Felsenstein, J. (1987). *PHYLIP (Phylogeny Inference Package) version 3.0 Manual.* University of Washington (distributed on floppy disks available from J. Felsenstein).

Hecht, M.K. & Edwards, J.L. (1976). The methodology of phylogenetic inference above the species level. In *Major Patterns in Vertebrate Evolution* (eds. M.K. Hecht, P.C. Goody, & B.M. Hecht). Plenum Press. New York.

Kluge, A.G. (1984). The relevance of parsimony to phylogenetic inference. In *Cladistics: Perspectives on the Reconstruction of Evolutionary History* (eds. T. Duncan & T.F. Stuessy). Columbia University Press. New York.

Moody, S.M. & O'Nolan, P. (1987). An *a priori* character weighting method for cladistic analysis. (Abstract) *American Zoologist* 26: 108A.

Neff, N.A. (1986). A rational basis for *a priori* character weighting. *Systematic Zoology* 35: 110-123.

Penny, D. & Hendy, M.D. (1985). Testing methods of evolutionary tree construction. *Cladistics* 1: 266-272.

Swofford, D.L. (1985). *PAUP: Phylogenetic Analysis Using Parsimony.* Illinois Natural History Survey. Illinois.

Watrous, L.E. and Wheeler, Q.D. (1981). The out-group comparison method of character analysis. *Systematic Zoology* 30: 1-11.